WHAT IS CLAIMED IS:

Sub A1

1. A method for segmenting an input sequence of characters from a non-segmented language, the method comprising:

identifying possible segments in the sequence of characters, at least two of the possible segments overlapping each other;

identifying an alternative sequence of characters for at least one of the possible segments, the alternative sequence of characters forming an alternative segment; and

performing multiple syntactic analyses using the possible segments and the alternative segment, the analyses resulting in a full syntactic parse that utilizes and thereby results in a segmentation of the input sequence of characters.

2. The method of claim 1 wherein performing multiple syntactic analyses comprises performing analyses that result in a parse containing the alternative segment.

3. The method of claim 1 wherein identifying an alternative sequence of characters for a possible segment comprises identifying an alternative sequence

of characters that has a different number of characters than the possible segment.

Sub A1

4. The method of claim 3 wherein performing multiple syntactic analyses comprises treating an alternative segment that has a different number of characters than the possible segment for which it is an alternative as if it had the same number of characters as the possible segment for which it is an alternative.

5. The method of claim 1 wherein identifying an alternative sequence of characters comprises performing inflectional morphology on a possible segment.

6. The method of claim 1 wherein identifying an alternative sequence of characters comprises identifying orthographic variations of a possible segment.

7. The method of claim 6 wherein identifying orthographic variations comprises identifying a preferred orthographic form for the possible segment.

8. The method of claim 1 wherein identifying orthographic variations comprises identifying orthographic variants that use a script other than the script of the characters in the possible segment.

Sub A1

~~9. A system for identifying syntax in a string of characters from a non-segmented language, the system comprising:~~

~~a word breaker that generates a collection of words from the string of characters, the collection of words comprising at least two words that are derived in part from the same character in the string of characters, the word breaker utilizing:~~

~~a lexical record set that is used to derive words for the collection of words by taking the words directly from the string of characters; and~~

~~a variants constructor that is used to derive word variants of words found in the string of characters, each word variant being added to the collection of words and having a different sequence of characters than the sequence of characters associated with the word in the string of characters from which it is derived; and~~

~~a syntax parser that performs a syntactic analysis using the collection of words produced by the word breaker to produce a syntax parse, the syntax parse~~

indicating the syntax of the string of characters.

Sub A1

10. The system of claim 9 wherein the variants constructor comprises inflectional morphology rules.

11. The system of claim 10 wherein the inflectional morphology rules are capable of identifying a word's lemma from its inflectional form in the string of characters.

12. The system of claim 9 wherein the variants constructor comprises an orthographic variants structure that indicates the orthographic variants of words found in the string of characters.

13. The system of claim 9 wherein at least one word variant has a different number of characters than the word from which it is derived.

14. The system of claim 9 wherein at least one word variant includes a character that is not present in the string of characters.

15. A computer-readable medium having computer-executable instructions for performing steps comprising:

Sub A1

receiving a sequence of characters that represent a phrase for a non-segmented language;

identifying a variant for a first group of characters in the sequence of characters, the variant containing a different collection of characters than the collection of characters in the first group of characters;

identifying a second group of characters in the sequence of characters that overlaps the first group of characters; and

performing a syntactic analysis using the variant and the second group of characters to produce a syntactic parse, the syntactic parse containing either the variant or the second group of characters.

16. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises identifying a variant that has a character that is not present in the first group of characters.

17. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises identifying a variant that has a

different number of characters than the first group of characters.

Sub A1

18. The computer-readable medium of claim 17 wherein identifying a variant comprises identifying a variant that has fewer characters than the first group of characters.

19. The computer-readable medium of claim 17 wherein identifying a variant comprises identifying a variant that has more characters than the first group of characters.

20. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises performing inflectional morphology on the first group of characters.

21. The computer-readable medium of claim 20 wherein the variant is a lemma of a word represented by the first group of characters.

22. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises identifying an orthographic variant of a word represented by the first group of characters.

Sub A1

~~23. The computer-readable medium of claim 22 wherein identifying an orthographic variant of a word comprises identifying a preferred orthographic form for a word.~~

~~24. The computer-readable medium of claim 22 wherein identifying an orthographic variant of a word comprises identifying a variant containing at least one character of a different script from the script of the characters in the first group of characters.~~